Provable Robust Watermarking for AI-Generated Text

Year: 2023 | Citations: 258 | Authors: Xuandong Zhao, P. Ananth, Lei Li, Yu-Xiang Wang

---

Abstract

We study the problem of watermarking large language models (LLMs) generated text -- one of the most promising approaches for addressing the safety challenges of LLM usage. In this paper, we propose a rigorous theoretical framework to quantify the effectiveness and robustness of LLM watermarks. We propose a robust and high-quality watermark method, Unigram-Watermark, by extending an existing approach with a simplified fixed grouping strategy. We prove that our watermark method enjoys guaranteed generation quality, correctness in watermark detection, and is robust against text editing and paraphrasing. Experiments on three varying LLMs and two datasets verify that our Unigram-Watermark achieves superior detection accuracy and comparable generation quality in perplexity, thus promoting the responsible use of LLMs. Code is available at https://github.com/XuandongZhao/Unigram-Watermark.